SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Cotton genome plus
• RESEARCH PAPER •

February 2016 Vol.59 No.2: 164–171
doi: 10.1007/s11427-016-5000-2

Transcriptome analysis reveals long noncoding RNAs involved in fiber development in cotton (*Gossypium arboreum*)

Changsong Zou, Qiaolian Wang, Cairui Lu, Wencui Yang, Youping Zhang, Hailiang Cheng, Xiaoxu Feng, Mtawa Andrew Prosper & Guoli Song*

State Key Laboratory of Cotton Biology, Institute of Cotton Research, Chinese Academy of Agricultural Sciences, Anyang 455000, China

Received May 14, 2015; accepted July 20, 2015; published online January 22, 2016

Long noncoding RNAs (lncRNAs) play important roles in various biological regulatory processes in yeast, mammals, and plants. However, no systematic identification of lncRNAs has been reported in *Gossypium arboreum*. In this study, the strand-specific RNA sequencing (ssRNA-seq) of samples from cotton fibers and leaves was performed, and lncRNAs involved in fiber initiation and elongation processes were systematically identified and analyzed. We identified 5,996 lncRNAs, of which 3,510 and 2,486 can be classified as long intergenic noncoding RNAs (lincRNAs) and natural antisense transcripts (lncNAT), respectively. LincRNAs and lncNATs are similar in many aspects, but have some differences in exon number, exon length, and transcript length. Expression analysis revealed that 51.9% of lincRNAs and 54.5% of lncNATs transcripts were preferentially expressed at one stage of fiber development, and were significantly highly expressed than protein-coding transcripts (21.7%). During the fiber and rapid elongation stages, rapid and dynamic changes in lncRNAs may contribute to fiber development in cotton. This work describes a set of lncRNAs that are involved in fiber development. The characterization and expression analysis of lncRNAs will facilitate future studies on their roles in fiber development in cotton.

long noncoding RNAs, strand specific RNA sequencing, fiber, transcriptome, expression

Citation: Zou, C., Wang, Q., Lu, C., Yang, W., Zhang, Y., Cheng, H., Feng, X., Prosper, M.A., and Song, G. (2016). Transcriptome analysis reveals long noncoding RNAs involved in fiber development in cotton (*Gossypium arboreum*). Sci China Life Sci 59, 164−171. doi: 10.1007/s11427-016-5000-2

INTRODUCTION

With the development of DNA sequencing technology and transcriptome analysis in recent years, the traditional view that protein-coding genes are the only effectors of gene function has been challenged. Long noncoding RNAs (lncRNAs) have been identified as a major component of the eukaryotic transcriptomes involved in the regulation of important biological processes (Derrien et al., 2012; Fabbri and Calin, 2010; Rinn and Chang, 2012; Zhang et al., 2010). Based on their relative positions with respect to protein-coding genes, those located in the intergenic regions were defined as long intergenic noncoding RNAs (lincRNAs), and long noncoding RNAs that partially overlapped with protein coding genes were referred to as natural antisense transcripts (lncNATs) (Derrien et al., 2012). Most lncRNAs, unlike protein-coding genes, lack sequence conservation between species in both plants and animals (Dong and Chen, 2013; Necsulea et al., 2014; Zhang et al., 2014). LincRNAs in higher eukaryotes might be transcribed by RNA polymerase II and processed by both 5′-capping and 3′-poly(A) additions (Guttman et al., 2009), and like most of the protein-coding genes, many contain one or more introns (Liu et al., 2012; Ulitsky et al., 2011; Zhu et al., 2013). LncRNAs are usually expressed at low levels and often exhibit tissue-specific patterns (Cabili et al., 2011; Rinn and Chang, 2012; Zhang et al., 2014), raising the possibility that lncRNAs participate in tissue development. LncRNAs, such

*Corresponding author (email: sglzms@163.com)

as Kcnq1ot1, bxd, and HOTAIR, are crucial for the precise control of embryogenesis in animals (Rinn et al., 2007; Umlauf et al., 2004). A recent study shows that lncRNAs may play an important role in *de novo* protein evolution (Ruiz-Orera et al., 2014). In plants, COLDAIR of lincRNA might participate in the epigenetic repression of *FLOWERING LOCUS C* (*FLC*) during vernalization (Heo and Sung, 2011), and one of rice lncRNAs has been identified to play a role in panicle development and fertility (Zhang et al., 2014). Long day specific male fertility associated RNA in rice was essential to normal pollen development under long-day conditions (Ding et al., 2012). In addition, a recent study showed that lncRNAs might play an important role in protein evolution (Ruiz-Orera et al., 2014).

With the rapid development in "omics" sequencing technology, lncRNAs have continued to be located in more plant species. In addition to the lncRNAs found in yeast and humans (Cabili et al., 2011; Ulitsky et al., 2011), more than 6,000 lincRNAs have been identified using a reproducibility-based bioinformatics strategy in *Arabidopsis* (Liu et al., 2012). Most recently, 35,268 lncNATs of *G. babardense* were identified and their expression patterns were characterized (Wang et al., 2015). In plant monocots, 2,224 lncRNAs were identified by strand specific RNA-sequencing in rice (Zhang et al., 2014), and 20,163 lincRNAs were identified by integration methods in maize (Li et al., 2014b).

Cotton (*Gossypium* spp.) is one of the most economically important crop plants with approximately 33 million ha planted per year worldwide. Its single-celled fiber is the principal natural source for the textile industry. *Gossypium* belongs to the Malvaceae family, and it diverged from a common ancestor with *Theobroma cacao* (Li et al., 2015; Li et al., 2014a; Wang et al., 2012). Based on the collective observations of pairing behavior, chromosome size, and relative fertility in interspecific hybrids, eight diploid subgenomes, designated as A to G and K, have been found across North America, Africa, Asia, and Australia. The haploid genome size of diploid cottons ($2n$=26) varies from about 880 Mb (*G. raimondii*) in the D genome to 2,500 Mb in the K genome (Hawkins et al., 2006; Hendrix and Stewart, 2005). The tetraploid cotton species ($2n$=4×=52), such as *G. hirsutum* and *G. barbadense*, are thought to have formed by an allopolyploidization event that occurred approximately 1–2 million years ago (Chen et al., 2007; Sunilkumar et al., 2006). Interestingly, the A genome species produce spinnable fibers that are cultivated on a limited scale, whereas the D genome species do not. Interestingly, the AD genome species can produce more suitable textile fibers than the A genome. Due to its excellent genetic and genomic resources, cotton is regarded as a good model to study genome polyploidization, and cotton fibers are an excellent experimental system for studying cell fate determination, cell elongation, and cell wall formation (Guan and Chen, 2013).

A previous study has completed the genome sequencing for cotton D, A, and AD (Cao, 2015; Li et al., 2015; Li et al., 2014a; Wang et al., 2012). Previous studies on noncoding RNAs in cotton have been largely limited to small RNAs. For example, the microRNAs (miRNAs) involved in sterile males and somatic embryogenesis have been identified in cotton (Gong et al., 2013; Wei et al., 2013), and there were 257 novel miRNAs that might relate to cotton fiber elongation (Xue et al., 2013). *MiR828* and *miR858* play roles in the regulation of fiber development in allotetraploid *G. hirsutum* (Guan et al., 2014). In this study, we aimed to identify lncRNAs in the allotetraploid cotton species *G. arboreum*, referring to the complete genome sequence and strand specific RNA sequencing. We used nine stranded transcriptomic sequences, representing the main stages of cotton fiber development, to identify lncRNAs. We systematically identified cotton lncRNAs (including lincRNAs and antisense lncRNAs) with a specific focus on the lncRNAs that were expressed during fiber development. The dynamic changes and fiber-specific expression in lncNATs and lincRNAs may contribute to ovule and fiber development in cotton.

RESULTS

A computational approach for the identification of lncRNAs in cotton

In order to systematically identify lncRNAs related to cotton fiber development, we performed whole transcriptome strand-specific RNA sequencing for immature ovules (1 day prior to anthesis, −1 DPA), fiber cell initials (on the day of anthesis, 0 DPA), young fiber-bearing ovules (1 day post-anthesis, 1 DPA), fiber (10, 15 DPA), and leaves in *G. arboreum.*

By integrating the lncRNA computational identification methods (Liu et al., 2012; Zhang et al., 2014), we made a cotton lncRNA pipeline based on strand specific RNA-seq data (Figure 1) using six whole transcriptome ssRNA-seq data sets. All RNA-seq datasets were first mapped to the whole genome of *G. arboreum* in order to reconstruct the cotton transcriptome. After filtering out infrequently expressed transcripts and transcripts that overlapped with transposable elements, we identified 43,732 transcription units, and 80.0% (32,097/40,134) of the annotated mRNAs genes could be recovered. The efficient recovery of annotated protein-coding genes indicated that the dataset used was suitable for the recovery of novel transcribed regions of the cotton genome.

We then evaluated the coding potential of the remaining transcripts and obtained novel expressed lncRNAs. We used the Coding Potential Calculator (CPC) to predict the coding potential of each transcript (Kong et al., 2007). All transcripts with CPC scores >0 were discarded. To guarantee the thorough elimination of protein-coding transcripts, we

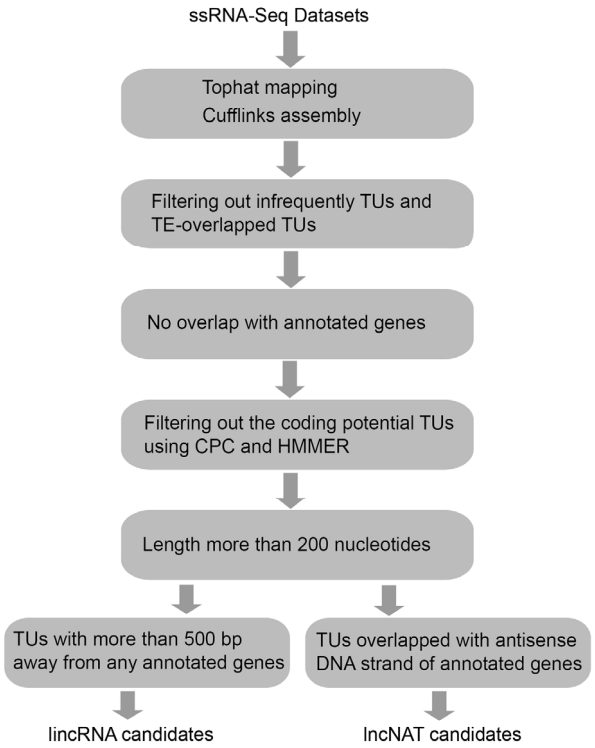

Figure 1 A computational pipeline for the systematic identification of lncRNAs in cotton. ssRNA-seq, strand specific RNA sequencing. TUs, transcription units. TE, transposable elements. CPC, coding potential calculator. HMMER, a software for biosequence analysis using profile hidden Markov models. lncNAT, long non-coding natural antisense transcript. ME, multiple exon. SE, single exon.

also employed HMMER (Eddy, 2009) to scan each transcript unit in all three reading frames to exclude transcripts that encoded any of the known protein domains cataloged in the Pfam protein family database (Punta et al., 2011). The following criteria were used to provide a strict definition for lncRNAs: (i) the transcript length must more than 200 nucleotides, and (ii) the transcript must contain no open reading frame (ORF) encoding more than 50 amino acids. Finally, the lncRNAs located at least 500 bp away from any annotated protein coding genes were defined as lincRNA, and the lncRNAs located on the antisense DNA strand and complementary to annotated genes were referred to as lncNAT. Ultimately, 3,510 lincRNA loci (Table S1 in Supporting Information) and 2,486 lncNAT loci (Table S2 in Supporting Information) were identified.

The characterization of lncRNAs in *G. arboreum*

To display the characteristics of lincRNAs and lncNATs more clearly, we analyzed the characteristics of lincRNAs and lncNATs separately in the following comparisons. We found that only a small fraction (median percentage, 10.8%) of the sequence for most of the lncNATs was antisense overlapped by protein-coding mRNA (Figure 2A) and that lincRNAs and lncNATs were similar in many aspects (Figure 2). The exon number distribution of lncRNAs showed that the *G. arboreum* genome encoded 74% of single-exonic lincRNAs and 72% of single-exonic lncNATs, which are significantly higher proportions than those of protein-coding transcripts (Figure 2B). Cotton lncRNAs have fewer exons

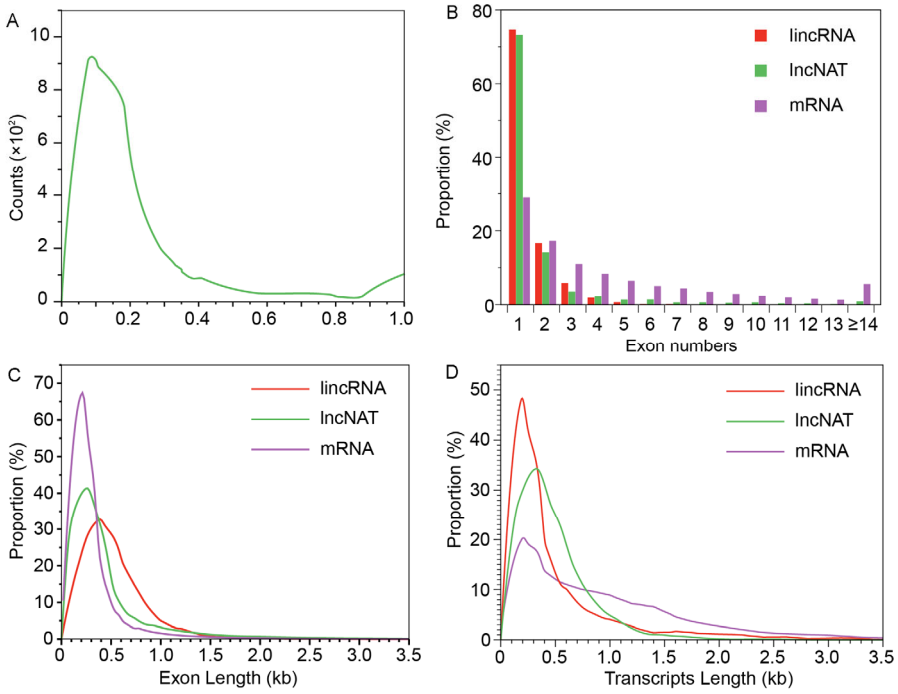

Figure 2 Properties of cotton lncRNAs. A, The proportion of lncNAT sequences overlapped by the annotated protein-coding genes. B, The number of exons per transcript for all lincRNAs and lncNATs and annotated protein-coding genes. C, Exon size distributions for lincRNAs, lncNATs and protein-coding transcripts. D, Transcript size distributions for lincRNAs, lncNATs and protein-coding transcripts.

than mRNAs (1.57 compared to 4.61 on average; 1.40 exons for lincRNAs and 1.82 exons for lncNATs), but their exon lengths (median length of 387 nucleotides, 422 nucleotides for lincRNAs, and 375 nucleotides for lncNATs) were significantly higher than those of mRNA (median length of 235 nucleotides) (Figure 2C). The mean transcript length of lncRNAs was typically lower than that of protein-coding genes (average lengths: 586 bp for lincRNAs, 629 bp for lncNATs and 1,088 bp for protein-coding transcripts) (Figure 2D), and about twice that of *Arabidopsis* (average length of 285 nucleotides) (Amor et al., 2009; Zhang et al., 2014). Like the *Arabidopsis* and rice lncRNA, only a small proportion of cotton lncRNAs (81 of 3,510 lincRNAs, 2.3%; 49 of 2,486 lncNATs, 2.0%) generate small regulatory RNAs (sRNAs) (Table S3 in Supporting Information), implying that these lncRNAs might function through generating sRNAs.

Expression characterization of cotton lncRNAs in fiber

The stranded RNA-seq data were used to systematically explore lncRNA expression among different fiber development stages. We estimated the overall expression level of each transcript using reads per kilobase of exon model per million (RPKM) and found that the lincRNAs and lncNATs were expressed at similar levels (means: 8.12 RPKM for lincRNA, 8.27 RPKM for lncNAT, respectively). These were significantly lower than the levels at which protein-coding genes are expressed (median: 24.3 RPKM, both $P<1\times10^{-18}$, *t*-test) but higher than the levels at which TE-related mRNAs are expressed (median: 4.3 RPKM, both $P<2.1\times10^{-16}$, *t*-test) (Figure 3A). This observation is consistent with a previous study (Cabili et al., 2011; Liu et al., 2012; Zhang et al., 2014). There were 1,092 and 515 lincRNAs with RPKM$\geqslant$0.5 at the fiber initiation and rapid elongation stages, respectively (Tables S4 and S5 in Supporting Information); and 955 and 466 lncNATswith RPKM $\geqslant$0.5 at the fiber initiation and rapid elongation stages, respectively (Tables S6 and S7 in Supporting Information). The distribution of lncRNA DEGs was almost identical to that of the protein coding genes (Figure 3B and C).

Based on the Jensen-Shannon (JS) score (Cabili et al., 2011), the degree of the differential expression of lincRNAs, lncNATs, mRNAs were estimated among fiber samples. We found that the distributions of lincRNAs and lncNATs were significantly different from protein-coding transcripts (Kolmogorov-Smirnov test; *P*-value<1.47×10^{-12}; Figure 3D). When a specificity JS score of 0.5 was used as a threshold, we found that 51.9% of lincRNAs and 54.5% of lncNATs transcripts were preferentially expressed at one stage of fiber development, much higher than protein-coding transcripts (21.7%) (Table 1). The differentially expressed genes (DEGs) between samples were also detected using the two-fold change criteria and false discovery rate (FDR) (corrected *P* value$\leqslant$0.001) between samples.

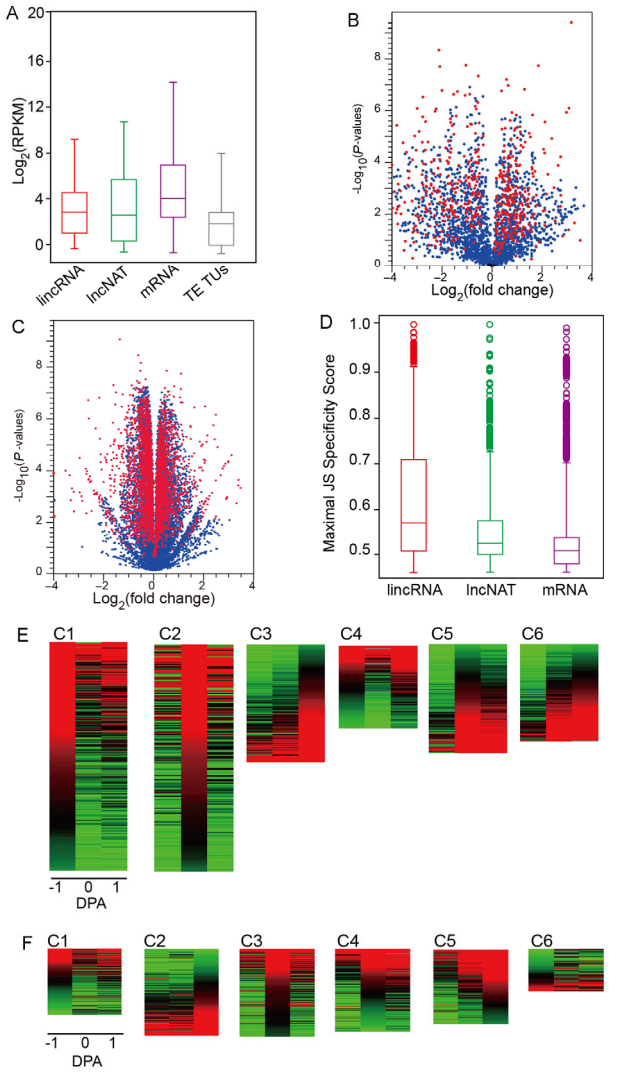

Figure 3 Characterization of cotton lncRNA expression. A, lncRNAs are transcribed at lower levels than protein coding genes but at higher levels than TE-mRNAs. RPKM, Reads Per Kilobase of exon model per Million. B, Volcano plot illustrating the distribution of fold changes and FDR *P* values for lincRNAs. The blue blots show the overall distribution, and the red blots represent the DEGs between 0 and 1 DPA samples. C, Volcano plot illustrating the distribution of fold changes and FDR *P* values for protein coding genes. The blue blots show the overall distribution, and the red blots represent the DEGs between 0 and 1 DPA samples. D, The distributions of maximal tissue specificity scores (JS scores) calculated for lncRNA and protein-coding transcripts of fiber and leaf tissues. E, *K*-means clustering of DEG lincRNAs from the −1, 0, and 1 DPA of fiber initiation, and the DEGs were grouped into six groups using the hierarchical clustering algorithm. F, *K*-means clustering of DEG lncNAT from the −1, 0, and 1 DPA of fiber initiation.

For example, there were 880 lincRNAs and 741 lncNATs that were identified as DEGs at the fiber initiation stage, and these genes were grouped into six clusters using *K*-means for clustering (Figure 3E and F), suggesting dramatic changes of lncNRAs during fiber development.

There were 19 randomly selected tissue-preferentially expressed lncRNAs verified by qRT-PCR (Figure 4). We found that it was in concordance with the results of

Table 1 Numbers of specific expression transcript units (TUs) for fiber and leaf samples.

TUs type	TUs numbers of JS scores (⩾0.5)						Percentage of total expressed TUs
	−1 DPA	0 DPA	1 DPA	10 DPA	15 DPA	Leaf	
lincRNA	203	274	258	162	107	461	51.9
lncNAT	145	219	174	151	123	234	54.5
mRNA	376	534	567	438	341	671	21.7

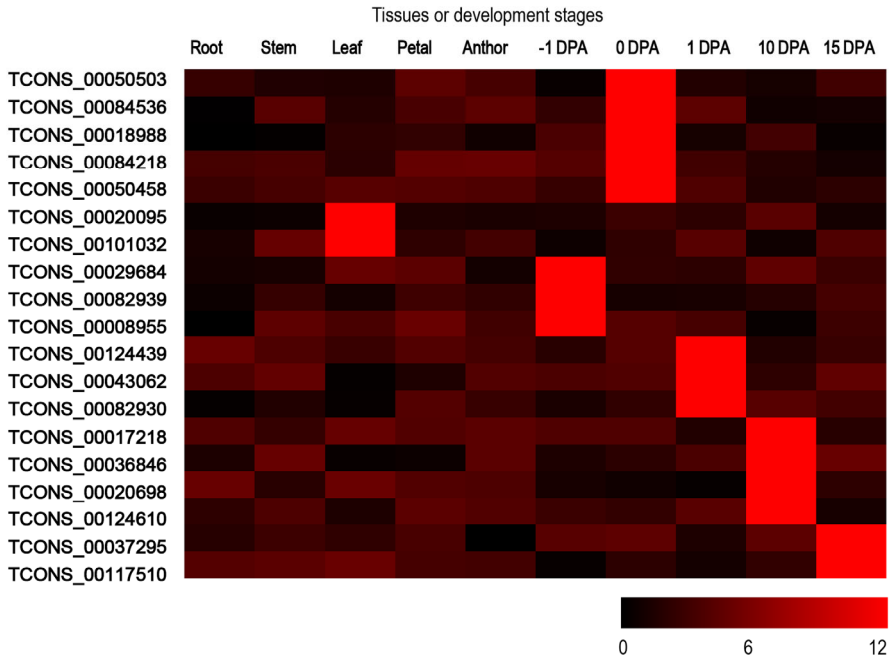

Figure 4 Expression of lncRNAs across 10 tissues or fiber developmental stages. −1, 0, and 1 DPA represents the ovules and fiber development stages. 5 and 10 DPA represent the fiber elongation development stages.

qRT-PCR and the ssRNA-seq results for most of the studied tissues, corroborating the reliability of lncRNA expression patterns based on ssRNA-seq data.

DISCUSSION

Fiber initiation and rapid elongation stages are crucial steps that affect the yield and quality of fiber, and are important for cotton because of its applications in agriculture. Over the past decade, genetic screens have identified a number of genes involved in fiber development (Pang et al., 2010; Qin et al., 2007; Qin and Zhu, 2011; Shi et al., 2006; Walford et al., 2011; Yan, 2015; Zou et al., 2013); however, the regulatory pathways that mediate the differentiation of the ovule epidermal and the fiber elongation process are far from being understood. Although an increasing number of reports indicate that lncRNAs function in the regulation of development inmammals and plants (Cabili et al., 2011; Liu et al., 2012; Ulitsky et al., 2011; Zhang et al., 2014), the identification of such lncRNAs in plants is in its infancy, and few plant lncRNAs have beenclearly identified to play roles in regulating plant development processes (Liu et al., 2012; Zhang et al., 2014). In this study, we systematically identified and analyzed cotton lncRNAs to find novel lncRNAs associated with fiber development. These data provide a good foundation for functional research of lncRNA in cotton fiber development.

Many pipelines have been reported in lncRNA identification (Liu et al., 2012; Wang et al., 2015; Zhang et al., 2014). In a previous study of *Arabidopsis*, the lncRNAs overlapping with TE were removed, and the lncRNAs were heavily methylated and kept silent (Fedoroff, 2012). Considering the previous study, the higher repetitive DNA sequence content (68.5%) of the genome (Li et al., 2014a), and the relatively low expression of TE-overlapped lncNRAs were filtered out of the transcripts located in the TE regions in this study. In addition, since protein domains vary in length from between about 25 amino acids up to 500 amino acids in length, we assumed that more strict standards should be used for the protein coding potential evaluation, and no ORF should encode more than 50 amino acids.

In summary, our current work identified a set of lncRNA involved in cotton fiber development using a bioinformatics approach, providing a platform for future investigation of cotton lncRNA regulation and function in fiber. Future work will aim to dissect their biological functions in relation to cotton fiber development and the genetics underpinning the improved fiber yield and quality. The *cis*-function

model had been recently reported with an intronic ncRNA named COLDAIR in plants (Heo and Sung, 2011). In cotton, gene expression is likely to be significantly regulated by diverse epigenetic modifications (Chen, 2007), and therefore, studies on lncRNAs are imperative, as some lncRNAs are probably involved in epigenetic regulation. We believe that cotton lncRNA have roles in dosage compensation, imprinting, enhancer function, and transcriptional regulation, and have a great impact on fiber development (Bonasio and Shiekhattar, 2014).

MATERIALS AND METHODS

Plant materials and samples

G. arboreum SXY1, derived from 18 successive generations of self-fertilization, was used for lncRNA analysis. The plants were cultivated in a field under normal conditions. Ovules and fibers were excised from developing flower buds or bolls on selected days post anthesis (DPA). Leaves were collected from two-week-old seedlings. The materials were quick-frozen in liquid nitrogen and stored at −70°C before use.

Transcriptome detection by RNA-seq

We sequenced strand specific RNA libraries from plant leaves derived from −1 DPA (ovules), 0 DPA (ovules), 1DPA (ovules), 10 DPA (fibers), 15 DPA (fibers), using illumina HiSeq2500 with 101-cycle pair-end sequencing protocols. Sequences were aligned to the whole genome of *G. arboreum* by using TopHat2 (Kim et al., 2013). The mapped sequences of each sample were assembled by Cufflinks version 2.1.1 with the annotation of *G. arboreum* as the reference (Li et al., 2014a). A non-redundant set of transcripts were calculated using cuffcompare (Trapnell et al., 2013).

Analysis for lincRNA identification and expression

We combined methods provided by Liu et al. and Zhang et al. to identify lncRNAs with minor adjustments (Liu et al., 2012; Zhang et al., 2014). The BEDtools (Quinlan, 2014) and in house perl scripts were also used to help the processes. The genomic loci of transcripts units (TUs) were compared with those of the annotated protein-coding genes and the annotations of TEs. Cleaned reads ($\geqslant$50 nucleotides in length after the quality check) were aligned with *G. arboreum* reference sequences using Bowtie2. Two mismatches were allowed per read. We then used Cufflinks to assess the expression level of each gene model using fragments per kilobase per million mapped read.

Quantitative reverse-transcription polymerase chain reaction (qRT-PCR)

A total of 1 to 2 mg of RNA previously treated with DNase I (NEB) was reverse transcribed using SuperScript III (TaKaRa) and lncRNA specific primer. cDNA was analyzed by quantitative PCR using SYBR Green Jump-Start Taq ReadyMix (Sigma-Aldrich) and the Applied Biosystems 7900HT real-time PCR system. All qRT-PCR reactions were performed in triplicates for each cDNA sample with an annealing temperature of 60°C for 40amplification cycles. Expression levels were quantified relative to that of the housekeeping gene *GaUBQ*7 (GenBank accession No. JZ555258). The comparative cycle threshold method was used to quantify relative expression levels of target transcripts. Primer sequences are presented in Table S1.

Compliance and ethics *The author(s) declare that they have no conflict of interest.*

Acknowledgements *The work was supported by the National Natural Science Foundation of China (31301369, 31271768, 31401425).*

Amor, B.B., Wirth, S., Merchan, F., Laporte, P., d'Aubenton-Carafa, Y., Hirsch, J., Maizel, A., Mallory, A., Lucas, A., and Deragon, J.M. (2009). Novel long non-protein coding RNAs involved in *Arabidopsis* differentiation and stress responses. Genome Res 19, 57–69.

Bonasio, R., and Shiekhattar, R. (2014). Regulation of transcription by long noncoding RNAs. Annu Rev Genet 48, 433.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25, 1915–1927.

Cao, X. (2015). Whole genome sequencing of cotton—a new chapter in cotton genomics. Sci China Life Sci 58, 515–516.

Chen, Z. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. Annu Rev Plant Biol 58, 377.

Chen, Z., Scheffler, B.E., Dennis, E., Triplett, B.A., Zhang, T., Guo, W., Chen, X., Stelly, D.M., Rabinowicz, P.D., and Town, C.D. (2007). Toward sequencing cotton (*Gossypium*) genomes. Plant Physiol 145, 1303–1310.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., and Knowles, D.G. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 22, 1775–1789.

Ding, J., Lu, Q., Ouyang, Y., Mao, H., Zhang, P., Yao, J., Xu, C., Li, X., Xiao, J., and Zhang, Q. (2012). A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. Proc Natl Acad Sci USA 109, 2654–2659.

Dong, Z., and Chen, Y. (2013). Transcriptomics: advances and approaches. Sci China Life Sci 56, 960–967.

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. Genome informatics. International Conference on Genome Informatics. 205–211.

Fabbri, M., and Calin, G.A. (2010). Beyond genomics: interpreting the 93% of the human genome that does not encode proteins. Curr Opin Drug Discov Dev 13, 350–358.

Fedoroff, N.V. (2012). Transposable elements, epigenetics, and genome evolution. Science 338, 758–767.

Gong, L., Kakrana, A., Arikit, S., Meyers, B.C., and Wendel, J.F. (2013). Composition and expression of conserved microRNA genes in diploid cotton (*Gossypium*) species. Genome Biol Evol 5, 2449–2459.

Guan, X., and Chen, Z. (2013). Cotton fiber genomics. Seed Genomics, 203–216.

Guan, X., Pang, M., Nah, G., Shi, X., Ye, W., Stelly, D.M., and Chen, Z. (2014). miR828 and miR858 regulate homoeologous *MYB2* gene

functions in *Arabidopsis* trichome and cotton fibre development. Nat Commun doi: 10.1038/ncomms4050

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., and Cassady, J.P. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458, 223–227.

Hawkins, J.S., Kim, H., Nason, J.D., Wing, R.A., and Wendel, J.F. (2006). Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome Res 16, 1252–1261.

Hendrix, B., and Stewart, J.M. (2005). Estimation of the nuclear DNA content of *Gossypium* species. Ann Bot 95, 789–797.

Heo, J.B., and Sung, S. (2011). Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science 331, 76–79.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36.

Kong, L., Zhang, Y., Ye, Z., Liu, X., Zhao, S., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 35, W345–W349.

Li, F., Fan, G., Lu, C., Xiao, G., Zou, C., Kohel, R.J., Ma, Z., Shang, H., Ma, X., and Wu, J. (2015). Genome sequence of cultivated Upland cotton (*Gossypium hirsutum* TM-1) provides insights into genome evolution. Nat Biotechnol 33, 524–530.

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., Li, Q., Ma, Z., Lu, C., and Zou, C. (2014a). Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat Genet 46, 567–572.

Li, L., Eichten, S.R., Shimizu, R., Petsch, K., Yeh, C.T., Wu, W., Chettoor, A.M., Givan, S.A., Cole, R.A., and Fowler, J.E. (2014b). Genome-wide discovery and characterization of maize long non-coding RNAs. Genome Biol 15, R40.

Liu, J., Jung, C., Xu, J., Wang, H., Deng, S., Bernad, L., Arenas-Huertero, C., and Chua, N. H. (2012). Genome-wide analysis uncovers regulation of long intergenic noncoding RNAs in *Arabidopsis*. Plant Cell 24, 4333–4345.

Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Gruetzner, F., and Kaessmann, H. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature 505, 635–640.

Pang, C.Y., Wang, H., Pang, Y., Xu, C., Jiao, Y., Qin, Y., Western, T.L., Yu, S., and Zhu, Y. (2010). Comparative proteomics indicates that biosynthesis of pectic precursors is important for cotton fiber and *Arabidopsis* root hair elongation. Mol Cell Proteomics 9, 2019–2033.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., and Clements, J. (2011). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 40, D290-D301.

Qin, Y., Hu, C., Pang, Y., Kastaniotis, A.J., Hiltunen, J.K., and Zhu, Y. (2007). Saturated very-long-chain fatty acids promote cotton fiber and *Arabidopsis* cell elongation by activating ethylene biosynthesis. Plant Cell 19, 3692–3704.

Qin, Y., and Zhu, Y. (2011). How cotton fibers elongate: a tale of linear cell-growth mode. Curr Opin Plant Biol 14, 106–111.

Quinlan, A.R. (2014). BEDTools: The swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics, doi: 10.1002/0471250953.bi1112s47.

Rinn, J.L., and Chang, H. (2012). Genome regulation by long noncoding RNAs. Annu Rev Biochem 81, 145–166.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., and Segal, E. (2007). Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. Cell 129, 1311–1323.

Ruiz-Orera, J., Messeguer, X., Subirana, J.A., and Alba, M.M. (2014). Long non-coding RNAs as a source of new peptides. Elife 3, e03523.

Shi, Y., Zhu, S., Mao, X., Feng, J., Qin, Y., Zhang, L., Cheng, J., Wei, L., Wang, Z., and Zhu, Y. (2006). Transcriptome profiling, molecular biological, and physiological studies reveal a major role for ethylene in cotton fiber cell elongation. Plant Cell 18, 651–664.

Sunilkumar, G., Campbell, L.M., Puckhaber, L., Stipanovic, R.D., and Rathore, K.S. (2006). Engineering cottonseed for use in human nutrition by tissue-specific reduction of toxic gossypol. Proc Natl Acad Sci USA 103, 18054–18059.

Trapnell, C., Hendrickson, D.G., Sauvageau, M., Goff, L., Rinn, J.L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31, 46–53.

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147, 1537–1550.

Umlauf, D., Goto, Y., Cao, R., Cerqueira, F., Wagschal, A., Zhang, Y., and Feil, R. (2004). Imprinting along the Kcnq1 domain on mouse chromosome 7 involves repressive histone methylation and recruitment of Polycomb group complexes. Nat Genet 36, 1296–1300.

Walford, S.A., Wu, Y., Llewellyn, D.J., and Dennis, E.S. (2011). GhMYB25-like: a key factor in early cotton fibre development. Plant J 65, 785–797.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., Yue, Z., Cong, L., Shang, H., and Zhu, S. (2012). The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet 44, 1098–1103.

Wang, M., Yuan, D., Tu, L., Gao, W., He, Y., Hu, H., Wang, P., Liu, N., Lindsey, K., and Zhang, X. (2015). Long noncoding RNAs and their proposed functions in fibre development of cotton (*Gossypium* spp.). New Phytol 207, 1181–1197

Wei, M., Wei, H., Wu, M., Song, M., Zhang, J., Yu, J., Fan, S., and Yu, S. (2013). Comparative expression profiling of miRNA during anther development in genetic male sterile and wild type cotton. BMC Plant Biol 13, 66.

Xue, W., Wang, Z., Du, M., Liu, Y., and Liu, J. (2013). Genome-wide analysis of small RNAs reveals eight fiber elongation-related and 257 novel microRNAs in elongating cotton fiber cells. BMC Genomics 14, 629.

Yan, P. (2015). The homeodomain-containing transcription factor, Gh HOX3, is a key regulator of cotton fiber elongation. Science 3, 013.

Zhang, Y., Liao, J., Li, Z., Yu, Y., Zhang, J., Li, Q., Qu, L., Shu, W., and Chen, Y. (2014). Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome Biol 15, 512.

Zhang, Y., Liu, J., Jia, C., Li, T., Wu, R., Wang, J., Chen, Y., Zou, X., Chen, R., and Wang, X.-J. (2010). Systematic identification and evolutionary features of rhesus monkey small nucleolar RNAs. BMC Genomics 11, 61.

Zhu, J., Fu, H., Wu, Y., and Zheng, X. (2013). Function of lncRNAs and approaches to lncRNA-protein interactions. Sci China Life Sci 56, 876–885.

Zou, C., Lu, C., Shang, H., Jing, X., Cheng, H., Zhang, Y., and Song, G. (2013). Genome-wide analysis of the *Sus* gene family in cotton. J Int Plant Biol 55, 643–653.

SUPPORTING INFORMATION

Table S1 The annotated GalincRNA gene of *G. arboretum.*

Table S2 The annotated GalncNAT gene of *G. arboreum.*

Table S3 Cotton lncRNAs could generate microRNA, and their most closed miRNA in miRBase.

Table S4 Expressed lincRNAs at fiber initiation stage.

Table S5 Expressed lincRNAs at fiber rapid elongation stage.

Table S6 Expressed lncNATs at fiber initiation stage.

Table S7 Expressed lncNATs at fiber rapid elongation stage.

Table S8 Relative expression level of lncRNA detected by RT-PCR.

Table S9 Primers used in this study.

The supporting information is available online at life.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.